Universidade de Brasília
Departamento de Estatística

Modelos de Regressão para Dados de Proporções Contínuas

Pedro Muniz Souza Silva

Relatório Final apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Pedro Muniz Souza Silva

Modelos de Regressão para Dados de Proporções Contínuas

Orientador:
Prof. Dr. **Leandro Tavares Correia**

Relatório Final apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Brasília
2018

Conteúdo

Resumo

Modelos de Regressão para Dados de Proporções Contínuas

Neste trabalho são abordados o modelo de regressão múltipla, modelo de regressão múltipla com transformação na variável resposta e o modelo de regressão beta. Os modelos foram aplicados em um conjunto de dados reais, obtidos a partir de fontes de informações públicas, referentes às eleições de 2018. Primeiro foram ajustados os modelos de regressão beta e foram comparados entre si, para definir qual o melhor ajuste. Numa segunda etapa foram ajustados modelos tradicionais de regressão (com pressuposto de normalidade). Por fim foram comparados os modelos, a fim verificar qual se ajusta melhor à natureza dos dados. Foi verificado que o modelo de regressão beta com dispersão variável tem vantagens em relação aos outros modelos testados.

Palavras-chave: Regressão Linear, Regressão Beta, Distribuição Beta, Eleições 2018.

1 Introdução

A modelagem para dados de proporções e taxas é uma parte importante na estatística que abrange diversas outras áreas. Um exemplo dessa aplicação pode ser visto em estudos demográficos quando se pretende analisar as taxas de crescimento populacional, taxa de fecundidade, a fim de propor políticas populacionais. No contexto da ciência agrônoma precisa-se estudar a proporção de um pesticida à ser aplicado em uma lavoura para que este não cause danos à plantação ou, em um caso mais grave, à população. Dito isso, este trabalho busca estudar modelos competitivos para analisar conjuntos de dados para dados de proporções e taxas.

Na literatura existem diversos estudos com modelos específicos para abordar essas estrutura de dados, pois modelos de regressão convencionais não são os mais adequados. Uma vez que esses conjuntos de dados não necessariamente irão atender as pressuposições do modelo de normalidade e homoscedasticidade, além de apresentar problemas na hora de restringir o domínio da variável resposta que, tratando-se de proporções e taxas, deve estar restrito ao intervalo $(0, 1)$. Para alterar o domínio da variável resposta usa-se alguns tipos de transformações e uma bem popular utilizada é a logito que altera o domínio de $(0, 1)$ para os $\mathbb{R}$. Contudo, certas transformações, como a logito, dificultam a interpretação dos parâmetros do modelo.

Uma alternativa à transformação do domínio da variável resposta é associar a variável resposta à uma distribuição que esteja definida no intervalo de interesse. Uma destas distribuições é a Beta, que tem sido considerada diversos trabalhos. Um dos trabalhos que vem sendo estudado é o modelo de regressão proposto por Ferrari and Cribari-Neto (2004) onde assume que a variável resposta segue uma distribuição Beta. Este modelo toma como base uma parametrização alternativa da densidade Beta em termos de média da variável resposta e um parâmetro de precisão.

A principal motivação para o estudo do modelo de regressão Beta está na flexibilidade da distribuição Beta. A densidade Beta assume várias formas diferentes dependendo da combinação dos valores dos parâmetros, incluindo esquerda e direita-enviesada ou a forma plana da densidade uniforme (que é um caso especial da densidade Beta mais geral). Deste modo, o modelo de regressão Beta acaba sendo mais flexível que os modelos sob suposição de normalidade.

Este trabalho pretende atingir dois objetivos: um relacionado a aspectos metodológicos e o outro a natureza empírica. O primeiro consiste em apresentar, descrever e caracterizar o modelo de regressão múltipla e o modelo de regressão beta, destacando aspectos inferenciais e de diagnósticos inerentes à análise de regressão. O segundo trata da aplicação e incorporação destes modelos de regressão para a estimação das proporções

de votos que os candidatos ao senado obtiveram nas eleições de 2018. Estes modelos serão comparados com a finalidade de encontrar o que melhor se ajusta à natureza dos dados.

2 Revisão de Literatura

2.1 Modelo de Regressão Linear Múltipla

Em diversas áreas da ciência busca-se explicar fenômenos experimentais ou observacionais, tais fenômenos podem ser explicados, ou até mesmo preditos, através de modelos de regressão. De forma geral uma regressão linear tem o objetivo de descrever uma variável resposta Y através de uma função linear de variáveis explicativas, isto é,

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_p x_{tp} + e_t; \quad t = 1, 2, ..., n, \tag{1}$$

em que Y_t representa o valor da t-ésima observação; x_{ti} representam os valores observados da t-ésima observação da variável i; $\beta_0, \beta_1, \beta_2, ..., \beta_p$ são os parâmetros da regressão; e e_t é o erro aleatório da t-ésima observação.

Além desse modelo, pode-se considerar um modelo um pouco mais complexo. Considere agora o seguinte modelo

$$Y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 (x_{t1} x_{t2}) + e_t \tag{2}$$

Neste caso, $(x_{t1}x_{t2})$ indica a interação entre x_{t1} e x_{t2}. Nos modelos em que existem interações e elas são significativas, o efeito de x_{t1} na resposta média depende do nível de x_{t2}, e vice-versa.

Assume-se para estes modelos que

- $E(e_t) = 0$, isto é, a média do erro é nula;
- $var(e_t) = \sigma^2$, isto é, o erro tem variância constante;
- $Cov(e_t, e_s) = 0, \forall t \neq s$, isto é, o erro de uma observação não é correlacionado com o erro de outra observação;
- $e_{t's} \sim N(0, \sigma^2)$, isto é, o erro segue uma distribuição Normal com média 0 e desvio padrão σ^2.

Deste modo, tem-se que

$$E(Y_i) \sim N(\beta_0 + \sum_{i=1}^{p} X_{it}\beta_i, \sigma^2) \tag{3}$$

2.1.1 Estimação por Mínimos Quadrados

O método consiste em minimizar a função

$$L = \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} \left(Y_t - \beta_0 - \beta_1 x_{t1} - \beta_2 x_{t2} - \ldots - \beta_p x_{tp}\right)^2 . \tag{4}$$

Derivando L em relação aos parâmetros $\beta' s$ obtêm-se

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{t=1}^{n} [Y_t - \beta_0 - \beta_1 x_{t1} - \beta_2 x_{t2} - \cdots - \beta_p x_{tp}],$$

$$\frac{\partial L}{\partial \beta_j} = -2 \sum_{t=1}^{n} [Y_t - \beta_0 - \beta_1 x_{t1} - \beta_2 x_{t2} - \cdots - \beta_p x_{tp}] x_{jt}, \quad j = 1, 2, \ldots, p$$

Para encontrar os pontos de mínimo, iguala-se as derivas à zero e substitui $\beta_0, \beta_1, \ldots, \beta_p$ por $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_p$, encontrando o seguinte sistema de equações normais

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{t1} + \widehat{\beta}_2 \sum_{t=1}^{n} x_{i2} + \ldots + \widehat{\beta}_p \sum_{t=1}^{n} x_{ip} = \sum_{t=1}^{n} Y_t \\ \widehat{\beta}_0 \sum_{t=1}^{n} x_{t1} + \widehat{\beta}_1 \sum_{t=1}^{n} x_{t1}^2 + \widehat{\beta}_2 \sum_{t=1}^{n} x_{t1}x_{t2} + \ldots + \widehat{\beta}_p \sum_{t=1}^{n} x_{t1}x_{tp} = \sum_{t=1}^{n} x_{t1}Y_t \\ \vdots\vdots \\ \widehat{\beta}_0 \sum_{t=1}^{n} x_{tp} + \widehat{\beta}_1 \sum_{t=1}^{n} x_{tp}x_{t1} + \widehat{\beta}_2 \sum_{t=1}^{n} x_{tp}x_{t2} + \ldots + \widehat{\beta}_p \sum_{t=1}^{n} x_{tp}^2 = \sum_{t=1}^{n} x_{tp}Y_t. \end{cases}$$

Resolvendo o sistema, obtêm-se os estimadores de mínimos quadrados $\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_p$.

Notemos que os estimadores de mínimos quadrados dos parâmetros do podem ser facilmente encontrados considerando a notação matricial dos dados, que é de fácil manipulação. Desta forma, o modelo de Regressão Linear Múltipla pode ser escrito como

$$Y = X\beta + e,$$

com

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{e} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix},$$

em que

- Y, é um vetor cujos componentes corresponde às n respostas;
- X, é uma matriz de dimensão $n \times (p+1)$ denominada matriz do modelo;
- β, é um vetor cujos elementos são os coeficientes de regressão;
- e, é um vetor de dimensão $n \times 1$ cujos componentes são os erros.

O método de mínimos quadrados tem como objetivo encontrar o vetor $\hat{\beta}$ que minimiza

$$L = \sum_{i=1}^{n} e_i^2 = e^\top e = (Y - X\beta)^\top (Y - X\beta) =$$

$$= Y^\top Y - Y^\top X\beta - \beta^\top X^\top Y + \beta^\top X^\top X\beta = Y^\top Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta,$$

,

sendo que $Y^\top X\beta = \beta^\top X^\top Y$ pois o produto resulta em um escalar. A notação $X^\top$ representa o transposto da matriz X enquanto que $Y^\top$ e $\beta^\top$ representam o transpostos dos vetores Y e β. Aplicando as derivadas parciais obtemos

$$\frac{\partial L}{\partial \beta} = -2X^\top Y + 2X^\top X\beta.$$

Igualando a zero e substituindo o vetor β por $\hat{\beta}$, obtemos

$$(X^\top X)\widehat{\beta} = X^\top Y.$$

Em geral, a matriz $(X^\top X)$ é não singular, ou seja, tem determinante diferente de zero, e portanto é invertível. Desta forma, conclui-se que os estimadores para os parâmetros $\beta_j, \quad j = 0, \dots, p$ são dados pelo vetor

$$\widehat{\beta} = (X^\top X)^{-1} X^\top Y. \tag{5}$$

A partir da Equação (5) podemos mostrar que o estimador $\hat{\beta}$ é não-viesado para β

da seguinte forma

$$\begin{aligned} E(\hat{\beta}) &= [(X^\top X)^{-1}X^\top]E(Y) \\ &= [(X^\top X)^{-1}X^\top]X\beta \\ &= [(X^\top X)^{-1}X^\top X]\beta \\ &= I\beta \\ &= \beta. \end{aligned}$$

Lembrando que a transposição de um produto é o produto de transpostas na ordem inversa (isto é, $(AB)^\top = B^\top A^\top$), que $X^\top X$ é simétrico, e que o inverso de uma transposição é a transposição do inverso, obtemos

$$\begin{aligned} Var(\hat{\beta}) &= [(X^\top X)^{-1}X^\top][Var(Y)][(X^\top X)^{-1}X^\top]^\top \\ &= [(X^\top X)^{-1}X^\top]I\sigma^2[(X^\top X)^{-1}X^\top]^\top \\ &= (X^\top X)^{-1}X^\top X(X^\top X)^{-1}\sigma^2 \\ &= (X^\top X)^{-1}\sigma^2 \end{aligned}$$

Assim, as variâncias e covariâncias dos coeficientes de regressão estimados são dadas pelos elementos de $(X^\top X)^{-1}$ multiplicado por σ^2. Os elementos da diagonal principal indicam as variâncias na ordem em que os coeficientes de regressão estão listados em β e os elementos fora da diagonal representam suas covariâncias, isto é, $Var(\hat{\beta_{ij}})$ representa a covariância entre i e j. Quando e segue uma distribuição Normal, $\hat{\beta}$ segue uma distribuição normal multivariada. Portanto

$$\hat{\beta} \sim N(\beta, (X^\top X)^{-1}\sigma^2).$$

2.2 Modelo de Regressão Múltipla Alternativo

Como visto na Equação (3), assume-se que os valores esperados da variável resposta seguem uma distribuição Normal onde o suporte são os reais ($\mathbb{R}'s$). Por isso ao se trabalhar com variáveis restritas a um intervalo, estes modelos acabam se tornando inadequados, visto que, podem ser preditos valores fora do suporte da variável resposta. Dessa maneira, uma alternativa é aplicar uma transformação que associe um número no intervalo $(0, 1)$ ao intervalo $(-\infty, \infty)$. Uma função que faz essa ligação é a logito.

2.2.1 Transformação Logito

Considere y definido no intervalo (0,1), a transformação logito é definida da seguinte forma

$$logito(y) = \nu = \log\left(\frac{y}{1-y}\right), \tag{6}$$

isto é,

$$logito : y \longrightarrow \nu,$$

de tal forma que $\{\nu \in \mathbb{R}\}$.

Aplicar esse tipo de transformação na variável resposta soluciona o problema do suporte da variável, porém o modelo ainda fica restrito aos pressupostos apresentados na Subseção 2.1. Contudo, esses tipos de transformações dificultam as interpretações dos parâmetros do modelo, no caso da transformação logito eles são interpretados em termos de ν.

2.3 Distribuição Beta

A distribuição Beta é uma distribuição contínua com dois parâmetros, α e β e está definida no intervalo $(0, 1)$. Sua função de densidade é dada por

$$f(y; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \tag{7}$$

$\forall y \in (0, 1)$, α>0 e β >0.

Como seu suporte está definido entre $(0, 1)$, ela acaba sendo frequentemente utilizada para modelar taxas, proporções, razões ou qualquer outro conjunto de dados contínuos pertencentes à este intervalo. Contudo, cabe ressaltar que a distribuição Beta não se restringe apenas ao intervalo (0,1), ela pode também ser modelada em um intervalo (a, b) por meio de uma transformação, onde a e b são escalares conhecidos e $a < b$. Outro ponto interessante sobre ela é que sua densidade pode assumir diferentes formas apenas mudando seus parâmetros α e β.

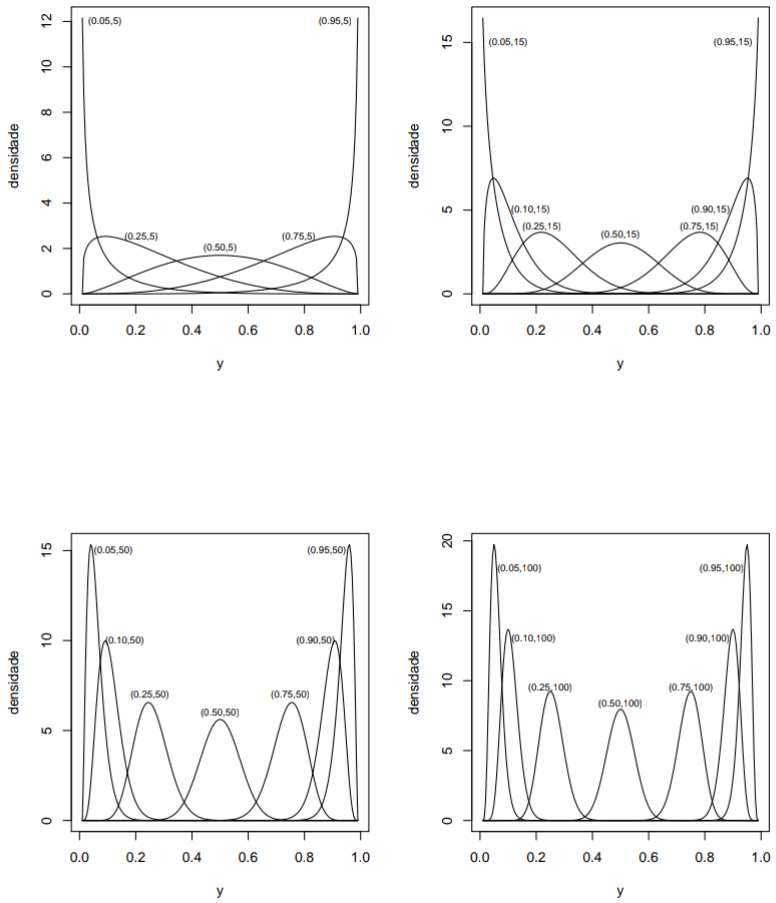

Figura 1: Densidade da Beta para diferentes combinações dos parâmetros

2.3.1 Função Geradora de Momentos

A Função Geradora de Momentos da distribuição Beta é dada por

$$\begin{aligned}
E(x^k) &= \int_{-\infty}^{\infty} x^k f(x;\alpha,\beta)dx \\
&= \int_0^1 x^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+k+\beta)} \int_0^1 \frac{\Gamma(\alpha+k+\beta)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}dx \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\beta)}{\Gamma(\alpha+k+\beta)}.
\end{aligned}$$

Usando a seguinte propriedade da função gama

$$\Gamma(z+1) = z\Gamma(z),$$

obtemos

$$E(x^k) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha)\alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k-1)\Gamma(\beta)}{\Gamma(\alpha+\beta)(\alpha+\beta)(\alpha+\beta+1)\cdots(\alpha+\beta+k-1)} = \prod_{i=0}^{k-1} \frac{\alpha+i}{\alpha+\beta+i}. \tag{8}$$

2.3.2 Esperança e Variância

A partir da Equação (8) é possível encontrar esperança e a variância de uma distribuição Beta

$$E(Y) = \frac{\alpha}{\alpha+\beta} \tag{9}$$

e, sabendo que $Var(Y) = E(Y^2) - [E(Y)]^2$, obtém-se

$$Var(Y) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \tag{10}$$

3 Metodologia

3.1 Modelo de Regressão Beta

Como vimos na Subseção 2.3, a distribuição beta é muito flexível e, portanto, é comumente utilizada para modelar dados restritos a um intervalo. As aplicações se tornam mais interessante quando o intervalo em questão é o intervalo unitário padrão, (0, 1). Pois, nestes casos, os dados podem ser interpretados como taxas ou proporções. Para trabalhar com essa distribuição, usando uma estrutura de regressão semelhante à classe conhecida de modelos lineares generalizados McCullagh and Nelder (1989), Ferrari and Cribari-Neto (2004) propuseram o Modelo de Regressão Beta.

Para o modelo de regressão proposto por Ferrari and Cribari-Neto (2004) foi utilizado uma reparametrização na densidade da distribuição Beta, dessa maneira, permite-se a modelagem da resposta média através de uma estrutura de regressão que envolve também um parâmetro de precisão. Considere $\phi = \alpha + \beta$ e $\mu = \frac{\alpha}{\alpha+\beta}$, isto é, $\alpha = \mu\phi$ e $\beta = (1-\mu)\phi$. Segue da Equação 9 e da Equação 10 que

$$E(Y) = \mu \qquad\qquad var(y) = \frac{V(\mu)}{1+\phi},$$

onde $V(\mu) = \mu(1-\mu)$. Sendo ϕ o parâmetro de precisão. Se fixarmos a esperança de $Y(\mu)$, quanto maior for o valor de ϕ, menor será a variância de y. Com isso, a densidade de Y (Equação 7) é reescrita com a seguinte parametrização:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{11}$$

em que $0 < \mu < 1$ e $\phi > 0$

Considerando $y_1, ..., y_n$ variáveis aleatórias independentes, com $t = 1, ..., n$. O modelo é obtido assumindo que a média de y_t pode ser escrita como

$$g_1(\mu_t) = \sum_{i=1}^{k} x_{ti}\beta_i = \eta_t \tag{12}$$

em que $\beta = (\beta_1, ..., \beta_k)^T$ é o vetor de parâmetros desconhecidos a ser estimado ($\beta \in \mathbb{R}^k$), $x_{t1}, ..., x_{tk}$ são observações de k variáveis independentes e η_t é o preditor linear. Por fim, $g_1(.)$, a função de ligação $g_1 : (0,1) \to \mathbb{R}$, é estritamente monótona e duas vezes diferenciável. Consequentemente, a resposta média é obtida aplicando a inversa da função de ligação, isto é, $\mu_t = g_1^{-1}(\eta_t)$, e $var(y_t) = \mu_t(1-\mu_t)/(1+\phi)$. Várias funções de ligações podem ser utilizadas, as mais utilizadas são

- Logito: $g_1(\mu_t) = log(\mu_t/(1-\mu_t))$;
- Probito: $g_1(\mu_t) = \phi^{-1}(\mu_t)$;
- Log-log: $g_1(\mu_t) = -log(-log(\mu_t))$;
- Complementar Log-log: $g_1(\mu_t) = log(-log(1-\mu_t))$.

Para o modelo proposto por Ferrari and Cribari-Neto (2004) considera-se que o parâmetro de precisão é constante ao longo das observações. Uma extensão deste modelo foi apresentada por Simas et al. (2010), aqui o parâmetro de precisão passa a depender de outras variáveis, e não mais fixo. Deste modo, pode-se modelar a precisão de forma similar à média. Mais especificamente, $y_t \sim Beta(\mu_t, \phi_t)$, para $t = 1, ..., n$ independente, com funções de ligação:

$$g_1(\mu_t) = \eta_{1t} = \mathbf{x_t}^\top \beta, \tag{13}$$

$$g_2(\phi_t) = \eta_{2t} = \mathbf{z_t}^\top \theta, \tag{14}$$

onde $\beta = (\beta_0, \beta_1, ..., \beta_p)^\top$ e $\theta = (\theta_0, \theta_1, ..., \theta_h)^\top$ são os vetores de parâmetros da regressão. Assim como $g_1(.)$, $g_2(.)$ é uma função de ligação, sendo $g_2 : (0, \infty) \to \mathbb{R}$, e também estritamente monótoma e duas vezes diferenciável. Para os modelos de precisão as funções de ligações mais utilizadas são

- $g_2(\phi_t) = log(\phi_t)$,
- $g_2(\phi_t) = \sqrt{\phi_t}$.

Para encontrar os estimadores dos vetores de parâmetros, β e ϕ, do modelo de regressão beta, utiliza-se o estimador de máxima verossimilhança da Equação (11), dada pela seguinte função de log-verossimilhança

$$l_t(\mu_t, \phi_t) = log\Gamma(\phi_t) - log\Gamma(\mu_t\phi_t) - \Gamma((1-\mu_t)\phi_t) + (\mu_t\phi_t - 1)log(y_t) + ((1-\mu_t)\phi_t - 1)log(1-y_t) \tag{15}$$

No entanto, como os estimadores de máxima verossimilhança de β e ϕ, não apresentam uma forma fechada, como no caso dos modelos vistos na Subseção 2.1, é necessário utilizar um método numérico de maximização da função de log-verossimilhança através de um algoritmo de otimização não linear. Para maiores detalhes inferenciais e expressões matriciais do vetor Score e da matriz de Informação de Fisher ver Simas et al. (2010).

3.2 Análise de Diagnóstico

A análise de diagnóstico é de fundamental importância na implementação de um modelo ajustado para verificar a existência de possíveis afastamentos das suposições do modelo. A metodologia de diagnóstico inicia-se com a análise dos resíduos para detectar pontos discrepantes e avaliar a adequabilidade da distribuição proposta para a variável resposta.

Na literatura existem diversos estudos a respeito de análise de resíduos para modelos de regressão McCullagh and Nelder (1989) propõe uma padronização dos componentes do desvio, onde se procura corrigir os efeitos de assimetria e curtose. Outra estudo foi apresentado por Atkinson (1985), onde é proposto a construção de uma banda de confiança para os resíduos de regressão linear, permitindo uma melhor análise a fim de verificar se os resíduos apresentam a distribuição esperada para o modelo. Para os modelos de regressão beta Ferrari and Cribari-Neto (2004), Espinheira et al. (2008a) e Rocha and Simas (2011) propõe resíduos padronizados, ponderados e de *score*.

Outro método conhecido de diagnóstico, é o de eliminar observações e medir o impacto da retirada desta observação nas estimativas dos parâmetros, para usa-se a medida chamada de Distância de Cook, Cook (1977), que também foi estendida para os modelos de regressão beta. Outro aspecto importante na análise de diagnóstico é a detecção de observações influentes.

Esta seção está organizada da seguinte forma. Primeiro são mostrados métodos para a análise de diagnóstico, explicado os resíduos que serão utilizado para o trabalho, o Leverage Generalizado e a Distância de Cook. E depois são apresentados métodos para seleção de modelos.

3.2.1 Resíduo Padronizado

Em sua publicação Ferrari and Cribari-Neto (2004) definiu o resíduo padronizado como

$$r_i = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{VAR}(y_t)}}. \tag{16}$$

Para o caso do Modelo de Regressão Beta, $\widehat{Var}(y_t) = \frac{\hat{\mu}_t(1-\hat{\mu}_t)}{1+\hat{\phi}_t}$.

3.2.2 Resíduo Ponderado Padronizado 2

Posteriormente Espinheira et al. (2008b) sugeriram a utilização de resíduos padronizados obtidos da convergência do processo iterativo (algoritmo de Escore Fisher) para estimação dos parâmetros de regressão, denominado-os de resíduos ponderados padronizados 1 e 2. O resíduo ponderado padronizado 2 é definido por

$$r_t^{ww} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{v_t(1-h_{tt})}}, \tag{17}$$

em que $v_i = \Psi'(\mu_i\phi) + \Psi'((1-\mu_t)\phi)$ e h_{tt} é o t-ésimo termo da matriz chapéu(para detalhes, Espinheira et al. (2008b)).

3.2.3 Resíduo de Score Modificado

Em outro estudo Rocha and Simas (2011) definiu um resíduo que leva em consideração toda a discrepância do modelo. Definido como

$$r_t^{ww} = \frac{y_t^* - \hat{\mu}_t^*}{\sqrt{v_t(1-\widehat{GL}_{tt})}} \tag{18}$$

em que $\widehat{GL}_{tt}$ é o t-ésimo da diagonal da matriz do Leverage Generalizado.

3.2.4 Leverage Generalizado

O Leverage é de suma importância na análise de influência em modelos de regressão. Usualmente, é medido por h_{tt} que são os elementos da diagonal da Matriz **H**, que é mais conhecida como matriz chapéu (ou matriz de projeção) e é usado para avaliar a importância individual de cada observação no próprio valor ajustado. Na regressão linear múltipla, por exemplo, é razoável utilizar o h_{tt} como uma medida de influência da t-ésima observação sobre o valor predito $\hat{y}_t$. Supondo que todos exerçam a mesma influência sobre os valores ajustados, pode-se esperar que h_{tt} esteja próximo de $\frac{tr(\mathbf{H})}{n} = \frac{p}{n}$, onde p é o número de parâmetros do modelo. Sugere-se destacar as observações para as quais se observa $h_{tt} > \frac{2p}{n}$, estas observações são conhecidas como pontos de alavanca.

(Ferrari and Cribari-Neto, 2004) obtiveram a expressão de forma fechada para o leverage generalizado ($GL(\beta)$) do modelo de regressão beta, para isso, eles consideraram ϕ conhecido, e assim encontraram

$$GL(\beta) = D_\beta(-\frac{\partial^2 l}{\partial\beta\partial\beta^\top})^{-1}\frac{\partial^2 l}{\partial\beta\partial y^\top} \tag{19}$$

em que $D_\beta = \partial\mu/\partial\beta^\top = \frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\beta^\top} = \mathbf{TX}$. Sendo $\mathbb{T} = daig(1/g'(\mu_1), ..., 1/g'(\mu_n))$. A expressão dada por $-\frac{\partial^2 l}{\partial\beta\partial\beta^\top}$ pode ser expressa da seguinte forma

$$-\frac{\partial^2 l}{\partial\beta\partial\beta^\top} = \phi\mathbf{X}^\top\mathbf{QX}$$

onde $\mathbf{Q} = diag(q_1, ..., q_n)$ com

$$q_t = [\phi\{\ \Psi'(\mu_t\phi) + \Psi'((1-\mu_t)\phi)\} + (y_t^* - \mu_t^*)\frac{g''(\mu_t)}{g'(\mu_t)}]\frac{1}{(g'(\mu_t))^2}$$

, t=1,...,n. E $\Psi'(.)$ é a função trigama. Adicionalmente temos que

$$\frac{\partial^2 l}{\partial\beta\partial y^\top} = \phi\mathbf{X}^\top\mathbf{TM}$$

onde $\mathbf{M} = diag(m_1, ..., m_n)$ com $m_t = \frac{1}{y_t(1-y_t)}$. E assim obtém-se a expressão final de GL(β) como sendo

$$GL(\beta) = \mathbf{TX}(\mathbf{X}^\top\mathbf{Q})^{-\mathbf{1}}\mathbf{X}^\top\mathbf{TM} \tag{20}$$

3.2.5 Distância de Cook

Encontrar observações que exercem um peso desproporcional no modelo é um tópico importante na análise de diagnóstico. A distância de Cook mede o impacto de uma dada observação sobre todos os $\beta's$ valores ajustados do modelo. A fim de obter uma medida semelhante a Distância de Cook para o modelo de regressão beta, Espinheira et al. (2008a) propôs utilizar o deslocamento da função de máxima verossimilhança da seguinte forma

$$LD_t = 2\{l_t(\hat{\beta}_t) - l_t(\hat{\beta}_{-t})\} \tag{21}$$

onde $l_t(\hat{\beta})$ e $l_t(\hat{\beta}_{-t})$, são, respectivamente, a função de log-verossimilhança avaliada na estimativa de máxima verossimilhança dos $\beta's$ dos dados completos e a função de log-verossimilhança avaliada na estimativa de máxima verossimilhança dos $\beta's$ obtida pela retirada da t-ésima observação dos dados.

3.3 Critérios de Seleção

3.3.1 Teste de Razão de Verossimilhança

A função de verossimilhança contém toda a informação relevante para fazer inferência acerca de um vetor de parâmetros de interesse. Portanto, dentre as técnicas para avaliação de modelos estatísticos, os testes baseados na função de verossimilhança são amplamente empregados.

Seja Θ o espaço paramétrico e o teste dado por:

$$H_0 : \theta = \theta^{(0)}; H_a : \theta \neq \theta^{(0)}$$

em $\theta^{(0)} \in \theta_0$, o espaço paramétrico restrito, $\theta_0 \subset \Theta$.

A razão de verossimilhança mensura o quanto as evidências estatísticas corroboram com valores para o vetor de parâmetros diferentes daqueles especificados em H_0, ou seja, valores grandes da razão de verossimilhanças rejeitam a hipótese nula. A estatística do teste da razão de verossimilhança pode ser expressa por

$$RV = 2[l(\hat{\theta}) - l(\theta^{(0)})]$$

em que $l(\theta)$ é o logaritmo da função de verossimilhança relativo ao vetor θ e $\hat{\theta}$ é o vetor estimador de máxima verossimilhança do parâmetro que indexa o modelo completo. Sob certas condições e sob H_0, RV se distribui, assintoticamente, como uma χ^2_r, sendo r o número de parâmetros testados.

3.3.2 Coeficiente de Determinação R^2_p

Após o ajuste do modelo, é importante realizar análises de diagnósticos com o proposito de atestar a qualidade do modelo estimado. Para isso Ferrari and Cribari-Neto (2004) introduziram o pseudo-$R^2(R^2_p)$, uma medida global da varição explicada. Esta medida foi definida como sendo o quadrado do coeficiente de correlação entre $\hat{\eta}$ e $g(y)$. Assim como no coeficiente de determinação da regressão linear (R^2), o pseudo-R^2 está definido no intervalo $[0, 1]$, e uma associação perfeita entre $\hat{\eta}$ e $g(y)$, e portanto, também entre $\hat{\mu}$ e y, temos um $R^2_p = 1$.

3.3.3 Critério de informação de Akaike (AIC)

Para a seleção de modelos, Akaike desenvolveu o critério de informação de Akaike, AIC, que é uma estimativa da informação K-L, baseado na maximização da função de log-verossimilhança acrescida de uma penalidade associada ao número de parâmetros do modelo como base para a seleção de modelos. Defini-se o AIC como:

$$AIC = 2p - 2ln(\hat{L})$$

em que $\hat{L}$ é a função de máxima verossimilhança do modelo e p é o número de parâmetros. Sugere-se que o modelo escolhido seja aquele que apresentar o menor valor de AIC dentre todos os modelos considerados para determinado problema.

4 Resultados e Discussões

4.1 Material

Para ilustrar a aplicação da regressão Beta, em dados reais, utilizou-se três conjuntos de dados obtidos na página na página do Tribunal Superior Eleitoral (TSE). Os dados são referente as Eleições de 2018, especificamente referente ao cargo de Senador(a). Para montar o conjunto de dados para este trabalho foram coletadas informações sobre os candidatos, quantidade de votos que receberam por município e, por fim, a receita que cada candidato tinha disponível durante as eleições.

Os arquivos disponibilizados pelo TSE contém informação de todos os candidatos para todos os cargos (Deputado Estadual, Deputado Distrital, Deputado Federal, Senador, Governador e Presidente), por isso foi necessário aplicar um filtro para coletar apenas informações dos candidatos ao Senado. Sobre os dados referente às receitas dos candidatos, é importante deixar claro que são provisórios (até o momento em que este relatória está sendo escrito). Deste modo, informo que os dados, para este estudo, são do dia 11 de novembro de 2018 e foram gerados às 00:27:51 (essas informações estão disponíveis no arquivo disponibilizado pelo TSE).

Outro fator importante é que estamos avaliando apenas os candidatos que tiveram o pedido de candidatura deferido pelo TSE, isto é, estão aptos a concorrer as eleições. Uma vez que, os votos são contabilizados de forma nominal, isto é, contém informação da quantidade de votos que cada candidato recebeu em uma determinada seção eleitoral de um determinado município, e alguns candidatos que tiveram a candidatura indeferida receberam votos, porém estes votos são contabilizados como votos nulos. Como estes votos não são contabilizados como votos validos, resolveu-se retirar esses candidatos.

A Tabela 1 contém uma breve descrição a respeito das variáveis abordadas neste trabalho. A fonte de informação para obtenção dos dados foi a página *http://www.tse.jus.br*. Nesta página foram coletadas informações sobre as eleições de 2018, buscamos coletar informações sobre os candidatos ao Senado.

Tabela 1: Descrição das variáveis

Variável	**Definição**
Região	Sigla da região geográfica de abrangência do candidato
UF	Sigla da Unidade da Federação de abrangência do candidato
Percentual de votos	Porcentagem de votos válidos que o candidato obteve
logito(percentual de votos)	Transformação logito da variável Percentual de votos
Receita total	Receita total do candidato que o candidato tem disponível para campanha
Receita partidária	Receita proveniente do partido
Receita financiamento coletivo	Receita proveniente de Financiamento Coletivo (modalidade de arrecadação de recursos para campanhas eleitorais por meio de sítios na internet, aplicativos eletrônicos e outros recursos similares)
Receita pessoa física	Receita proveniente de doações de Pessoas Físicas
Receita própria	Receita proveniente do próprio candidato
Gênero	Descrição do gênero do candidato (masculino ou feminino)
Reeleição	Se o candidato está tentando a reeleição
Quantidade de partidos coligação	Quantidade de partidos na coligação do candidato

4.2 Descrição dos Dados

Inicialmente foi realizado uma análise exploratório dos dados para observar possíveis características das covariáveis. A Tabela 2 apresenta algumas estatísticas descritivas, como mínimo, primeiro quartil ($Q_{1/4}$), mediana, média, terceiro quartil ($Q_{3/4}$) e máximo das variáveis que serão utilizadas na modelagem da regressões. Com essas informações podemos tirar algumas conclusões. Por exemplo, o percentual de votos válidos obtidos pelos candidatos variam de 0.06047% até 46.26%. Sobre as variáveis relacionadas às receitas, achou-se melhor trabalhar usando a receita total e não a receita subdividida em categorias como receita partidária, receita de pessoas físicas, receita de financiamento coletivo e receita própria.

Tabela 2: Análise descritiva das variáveis

Variável	**Mínimo**	$Q_{1/4}$	**Mediana**	**Média**	$Q_{3/4}$	**Máximo**
PCT_VAL	0.0006047	0.0132635	0.0625729	0.0937557	0.1536902	0.4625842
TR_LOGIT	-7.4101	-4.3094	-2.7068	-3.0835	-1.7059	-0.1499
RCT_TOTAL	200	23904	383057	957906	1829060	5327774
RCT_FIN_COL	0	0	0	3167	0	267210
RCT_PF	0	0	5550	91803	62800	1481000
RCT_PROPRIO	0	0	1000	136278	23853	3377000
QT_PARTIDO_COL	1	1	3	4.754	7	18

Nota: PCT_VAL é o percentual de votos válidos; TR_LOGIT é a transformação logito de PCT_VAL; RCT_TOTAL é a receita total; RCT_FIN_COL é a receita oriunda de financiamento coletivo; RCT_PF é a receita oriunda de pessoas físicas; RCT_PROPRIA é a receita oriunda do próprio candidato; QT_PARTIDO_COL é a quantidade de partidos na coligação.

Na Figura 2 podemos notar uma grande diferença na distribuição de recita pelos partidos entre os candidatos que estão tentando a reeleição e os candidatos que estão concorrendo pela primeira vez ao cargo de Senador (temos 32 candidatos tentando a reeleição e 253 disputando um primeiro mandato ao cargo de senador). A situação já não é similar se observarmos as receitas dos candidatos por sexo (temos 52 mulheres concorrendo ao cargo de senadora contra 233 homens concorrendo ao mesmo cargo), podemos perceber na Figura 3 uma certa semelhança entre as receitas. Cabe ressaltar que, a Lei nº 13.488, de 6 de outubro de 2017 fixou o limite de gasto de campanha eleitoral em valores absolutos por cargo eletivo para as eleições de 2018. Para o cargo de Senador da República, o limite de gasto para a campanha é fixado de acordo com o eleitorado do Estado em 31 de maio de 2018, nos termos do Art. 5º da Resolução TSE nº 23.553/2017. Isto é, o limite de receita dos candidatos ao Senado é proporcional à quantidade de eleitores da Unidade da Federação para o qual o candidato está concorrendo.

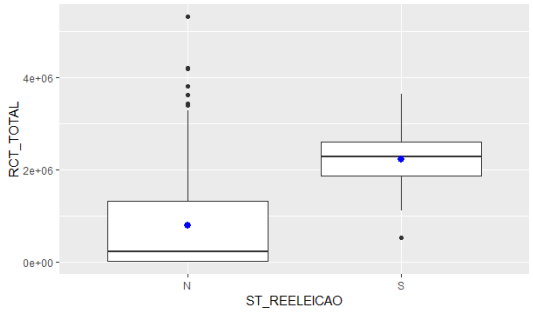

Figura 2: Boxplot da percentual de votos segundo situação de reeleição

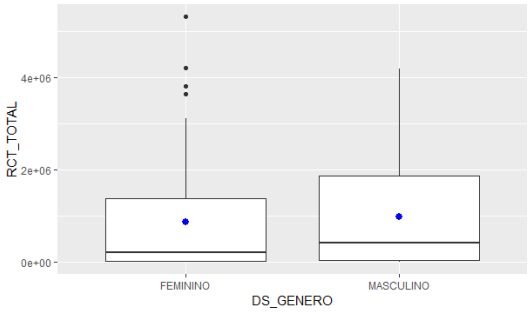

Figura 3: Boxplot da receita dos candidatos segundo sexo

Com o auxilio da Figura 4 podemos observar o comportamento da variável resposta que usaremos, porcentagem de votos válidos obtidos. Nela observamos um comportamento assimétrico à direita. Isso já nos mostra que ajustar o modelo com base nos pressupostos apresentados na Subseção 2.1, não seria o recomendado.

Na Figura 5 também é possível perceber o comportamento do percentual de votos com a variável Receita total. Note que os candidatos com pouca receita tendem a ter o percentual de votos concentrados em 0.1, e a medida que a receita disponível aumenta a variabilidade do percentual de votos também cresce, podendo ser um indício de usar a variável receita total como covariável no modelo de precisão. Neste gráfico quanto maior for o círculo que representa o candidato, maior será a quantidade de partidos na coligação do seu partido.

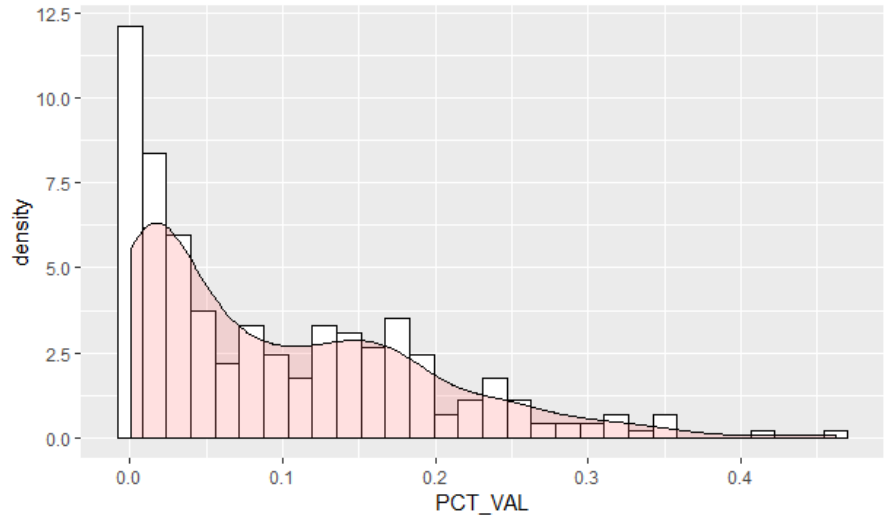

Figura 4: Histograma e Densidade da variável Percentual de Votos Válidos

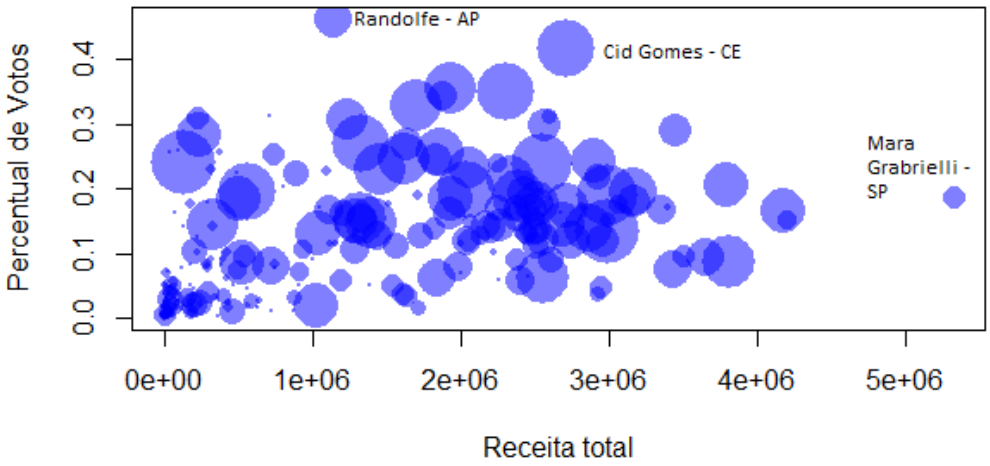

Figura 5: Gráfico de dispersão, Percentual de votos válidos por Receita total

Como a amplitude da receita dos candidatos é muito grande, sugere-se que seja usado o log(Receita total). Na Figura 6, percebemos a vantagem de se usar essa transformação, pois, podemos notar uma linearização entre as variáveis percentual de votos e log(receita total).

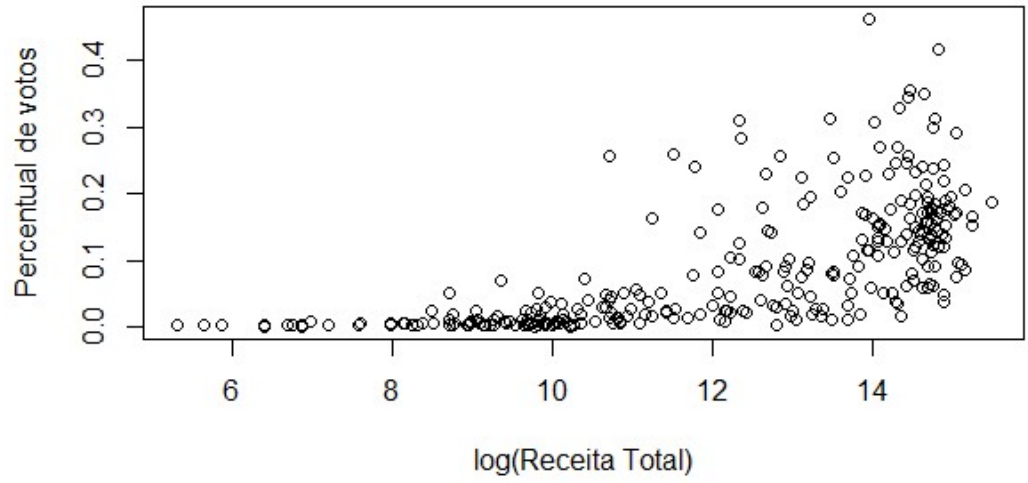

Figura 6: Gráfico de dispersão, Percentual de votos válidos por log(Receita total)

Tabela 3: Correlção entre as variáveis

	Percentual de Votos	logit(Percent. de Votos)	Receita Total	Receita Financ. Coletivo	Receita Pessoa Física	Receita Própria	Quantidade de Partidos na Coligação	Receita Partidária	Log(Receita Total
Percentual de Votos	1.00	0.86	0.58	0.03	0.32	0.23	0.62	0.50	0.65
logit(Percent. de Votos)	-	1.00	0.64	0.06	0.32	0.25	0.59	0.56	0.82
Receita Total	-	-	1.00	0.00	0.43	0.42	0.61	0.88	0.80
Receita Financ. Coletivo	-	-	-	1.00	0.03	-0.05	-0.02	-0.00	0.07
Receita Pessoa Física	-	-	-	-	1.00	0.19	0.39	0.17	0.37
Receita Própria	-	-	-	-	-	1.00	0.16	-0.01	0.32
Quantidade de Paridos na Coligação	-	-	-	-	-	-	1.00	0.55	0.59
Receita Partidária	-	-	-	-	-	-	-	1.00	0.71
Log(Receita Total)	-	-	-	-	-	-	-	-	1.00

Na Tabela 3 é apresentado a correlação de Pearson entre as variáveis. Nota-se que a receita que apresenta maior correlação entre as variáveis é a Receita total, sendo que o log(Receita total), apresenta um coeficiente ainda maior.

4.3 Resultados

Inicialmente, ao se ajustar o modelo de regressão beta, é importante se questionar a respeito da precisão dos dados. Pois, como já visto, modelos de regressão com dispersão variável precisam de uma estrutura para modelar a precisão dos parâmetros de modo a melhorar os resultados.

Por este motivo, para se ajustar o modelo que melhor explique a natureza dos dados, foi inicialmente considerado um modelo com todas as variáveis de estudo, considerando o parâmetro de precisão fixo. Em seguida, foi ajustado um modelo com o parâmetro de precisão variável, também com todas as variáveis, usando apenas a variável log(Receita total) para modelar a precisão, pois, como visto na Figura 6, há indícios de uma tendência linear na dispersão do percentual de votos com log(receita total). Ao final foi realizado um teste de razão de verossimilhança entre os dois modelos e comparado seus AIC's, a fim de verificar se o modelo com precisão variável apresenta melhorias em relação ao de precisão fixa.

Na Tabela 4 é apresentado o resultado do Teste da Razão de Verosimilhança entre os dois modelos. Verifica-se que há uma melhora ao se ajustar o modelo com precisão variável em relação ao de precisão fixa. Note que no teste de razão de verossimilhança estamos testando $H_0 : y_t \sim Beta(\mu_t; \phi)$ e $H_1 : y_t \sim Beta(\mu_t; \phi_t)$. Deste modo, como foi verificado um p-valor de 0.0007, rejeita-se H_0. Verifica-se ainda que o AIC do modelo com precisão variável é menor do que o com precisão fixa. Portanto é necessário uma estrutura de regressão para modelar a precisão dos dados.

Tabela 4: AIC e Teste da Razão de Verossimilhança

	AIC	LogLik	Df	χ^2	Pr($>\chi^2$)
Modelo 1	-1141.933	576.97			
Modelo 2	-1151.547	582.77	1	11.61	0.0007

O segundo passo foi encontrar um modelo que melhor se ajuste a natureza dos dados. Foi necessário primeiro criar um modelo para μ para, em seguida, ajustar um modelo para ϕ.

Para ajustar um modelo para μ, foi utilizado o seguinte método. Primeiro criou-se um modelo com todas as variáveis (considerando o modelo de precisão apenas com a variável log(receita total)). Foram mantidas no modelo aquelas variáveis que de fato apresentavam um impacto relevante na variável resposta. Em seguida, ainda foram criados mais dois modelos com interação, a fim de verificar a suposição da Subseção 4.2, que exitem interações entre as variáveis receita total e reeleição, e receita total e quantidade de partidos na coligação. Para esta análise foi considerado a precisão apenas com a

variável log(receita total). Depois será ajustado um modelo, de maneira semelhante, para ϕ.

Na Tabela 5, encontram-se as etapas utilizadas para encontrar o melhor modelo para μ. Observa-se que, para o primeiro modelo a única variável não significativa foi a reeleição. Então foi retirada esta variável, chegando ao segundo modelo, em que todas as variáveis deram significativas. Já para os modelos com interações, apenas a interação entre receita total e quantidade de partidos na coligação foi significativa. Por fim, o modelo escolhido para μ foi o modelo 2, por ser mais parcimonioso, e seu R_p^2(valor) está bem próximo ao do modelo 4, que teve maior R_p^2(valor) e menor AIC.

Para ajustar os modelos de ϕ, o método utilizado foi semelhante ao μ, porém a única interação testada foi entre log(receita total) e quantidade de partidos na coligação. Observe que, na Tabela 6, o primeiro modelo (modelo completo), já é possível perceber que a única variável significativa é a log(receita total). No caso do modelo 6, nota-se que a interação entre log(Receita total) e quantidade de partidos na coligação é significativa. Porém, o modelo escolhido, ficou sendo o modelo 2, pois, é o com menor AIC, e se trata de um modelo mais parcimonioso, com interpretação mais direta.

Tabela 5: Ajuste para μ dos Modelos de Regressão Beta

	Variável Explicativa	**Parâmetro**	**Estimativa**	**Erro Padrão**	**Pr(>|z|)**
Modelo 1	Intercepto	β_0	-7.50	0.35	0.00
	log(Receita total)	β_1	0.35	0.03	0.00
	Qt de partidos na coligação	β_2	0.06	0.01	0.00
	Gênero: Masculino	β_3	0.27	0.11	0.02
	Reeleição	β_4	0.01	0.12	0.93
	Intercepto	θ_0	69.39	15.30	0.00
	log(Receita total)	θ_1	-3.57	1.11	0.00
Modelo 2	Intercepto	β_0	-7.50	0.34	0.00
	log(Receita total)	β_1	0.36	0.03	0.00
	Qt de partidos na coligação	β_2	0.06	0.01	0.00
	Gênero: Masculino	β_3	0.27	0.11	0.02
	Intercepto	θ_0	69.41	15.30	0.00
	log(Receita total)	θ_1	-3.58	1.11	0.00
Modelo 3	Intercepto	β_0	-8.13	0.43	0.00
	log(Receita total)	β_1	0.40	0.03	0.00
	Qt de partidos na coligação	β_2	0.31	0.10	0.00
	Gênero: Masculino	β_3	0.27	0.11	0.02
	Reeleição	β_4	0.16	2.89	0.96
	Interação(log(Receita Total), Qt de partidos na coligação)	β_5	-0.02	0.01	0.01
	Interação(log(Receita Total), Reeleição)	β_6	-0.01	0.20	0.97
	Intercepto	θ_0	73.32	16.11	0.00
	log(Receita total)	θ_1	-3.81	1.17	0.00
Modelo 4	Intercepto	β_0	-8.16	0.43	0.00
	log(Receita total)	β_1	0.40	0.03	0.00
	Qt de partidos na coligação	β_2	0.31	0.10	0.00
	Gênero: Masculino	β_3	0.27	0.11	0.02
	Interação(log(Receita Total), Qt de partidos na coligação)	β_4	-0.02	0.01	0.01
	Intercepto	θ_0	73.41	16.12	0.00
	log(Receita total)	θ_1	-3.82	1.17	0.00

Tabela 6: Ajuste para ϕ dos Modelos de Regressão Beta

	Variável Explicativa	**Parâmetro**	**Estimativa**	**Erro Padrão**	**Pr(>|z|)**
Modelo 5	Intercepto	β_0	-7.54	0.35	0.00
	log(Receita total)	β_1	0.36	0.03	0.00
	Qt de partidos na coligação	β_2	0.05	0.01	0.00
	Gênero: Masculino	β_3	0.30	0.13	0.02
	Intercepto	ϕ_0	75.45	17.25	0.00
	log(Receita total)	ϕ_1	-4.01	1.30	0.00
	Qt de partidos na coligação	ϕ_2	0.19	0.47	0.68
	Reeleição	ϕ_3	5.51	5.54	0.32
	Gênero: Masculino	ϕ_4	-2.52	5.41	0.64
Modelo 2	Intercepto	β_0	-7.50	0.34	0.00
	log(Receita total)	β_1	0.36	0.03	0.00
	Qt de partidos na coligação	β_2	0.06	0.01	0.00
	Gênero: Masculino	β_3	0.27	0.11	0.02
	Intercepto	ϕ_0	69.41	15.30	0.00
	log(Receita total)	ϕ_1	-3.58	1.11	0.00
Modelo 6	Intercepto	β_0	-7.66	0.35	0.00
	log(Receita total)	β_1	0.37	0.03	0.00
	Qt de partidos na coligação	β_2	0.05	0.01	0.00
	Gênero: Masculino	β_3	0.26	0.11	0.02
	Intercepto	ϕ_0	95.64	21.22	0.00
	log(Receita total)	ϕ_1	-5.60	1.62	0.00
	Qt de partidos na coligação	ϕ_2	-8.20	3.56	0.02
	Interação(log(Receita total, Qt de partidos na coligação)	ϕ_3	0.60	0.26	0.02

Tabela 7: R_p^2 e AIC dos modelos ajustados

	R_p^2	**AIC**
Modelo 1	0.695	-1151.547
Modelo 2	0.695	-1153.541
Modelo 3	0.701	-1153.353
Modelo 4	0.701	-1157.258
Modelo 5	0.695	-1148.923
Modelo 6	0.696	-1152.638

4.3.1 Propostas de Funções de Ligação

Nos modelos de regressão beta há ainda a especificação da função de ligação tanto para μ, quanto para ϕ. Onde a escolha de uma determinada função de ligação pode melhorar significativamente o ajuste, especialmente quando se tem dados próximos aos valores extremos 0 e 1.

Com o objetivo de comparar o ajuste, utilizando as funções de ligação para μ e para ϕ, apresentadas na Subseção 3.1, a Tabela 8 apresenta os AIC's para diferente tipos de função de ligação. Observa-se que o melhor modelo, pelo critério AIC, foi o com função de ligação Probito para μ e função de ligação *log* para ϕ. Porém a diferença entre os modelos não está tão acentuada.

Já na Tabela 9 é apresentado os valores de R_p^2 para os diferentes tipos de função de ligação. Nota-se que, com exceção dos modelos em que foi usado a função de ligação *Loglog* para μ, os valores dos coeficientes de determinação estão bem próximos.

Portanto, como não foi evidenciada nenhuma melhora expressiva entre as diferentes funções de ligação, achou melhor trabalhar com o modelo que usa *Logito* como função de ligação para e μ e *Log* como função de ligação para ϕ. Uma vez que, está entre os melhores R_p^2 e AIC.

Tabela 8: AIC para diferentes Funções de Ligação

$\phi \setminus \mu$	**Logito**	**Probito**	**Cloglog**	**Loglog**
Identidade	-1153.541	-1152.597	-1152.468	-1148.148
Log	-1164.264	-1165.212	-1162.513	-1160.415
Raíz Quadrada	-1157.049	-1156.548	-1155.779	-1151.924

Tabela 9: R_p^2 para diferentes Funções de Ligação

$\phi \setminus \mu$	**Logito**	**Probito**	**Cloglog**	**Loglog**
Identidade	0.695	0.686	0.700	0.664
Log	0.697	0.688	0.699	0.665
Raíz Quadrada	0.696	0.687	0.698	0.664

Tabela 10: Modelo final escolhido

Modelo para μ, função de ligação Logito			
Variáveis	$\hat{\beta}$	Erro padrão	P-valor
Intercepto	-8.21	0.33	<2e-16
Log(Receita)	0.41	0.03	<2e-16
Qt partidos coligação	0.05	0.01	<e-07
Gênero:Masculino	0.25	0.11	0.03
Modelo para ϕ, função de ligação Log			
Variáveis	$\hat{\theta}$	Erro padrão	P-valor
Intercepto	6.46947	0.45974	<2e-16
Log(Receita)	-0.26174	0.03622	<e-13

Agora, para interpretação das estimativas dos parâmetros, será utilizado a razão de chances, através da média $e^{c\times\hat{\beta}_i}$, onde c é o incremento na variável. Deste modo, se houver um incremento de 0.1 na variável log(Receita total), temos um acréscimo de 4.19% na média da variável resposta. No caso de se aumentar 1 partido na coligação, o candidato terá um acréscimo de 5.13% na média da variável resposta.

Analisando agora o modelo de precisão, nota-se que a receita contribui também para diminuir a precisão do modelo, pois um incremento de 0.1 no log(receita total) reduz em 2.58% a precisão dos dados.

4.3.2 Análise de Diagnóstico do Modelo Ajustado

A análise de diagnóstico se apresenta como uma importante etapa na construção dos modelos de regressão, pois é nela que se investiga algumas medidas da qualidade do ajuste além das suposições feitas sobre o modelo, tais como adequação da distribuição de probabilidade suposta para a variável resposta, aleatoriedade dos resíduos e análise das medidas de influência. Para a análise do modelo ajustado será utilizado os resíduos ponderados padronizados proposto por Espinheira et al. (2008b), uma vez que, os resíduos de score modificado, proposto por Rocha and Simas (2011), não apresentaram uma melhora significativas nas análises.

O gráfico de probabilidade normal com envelope simulado é uma técnica utilizada para analisar possíveis desvios da suposição do modelo e observações discrepantes. Na Figura 7 verifica-se que as observações encontram-se, na sua maioria, dentro dos limites do envelope. Portanto, não há evidências suficientes para descordar da adequabilidade do modelo. Por outro lado, a Figura 8 apresenta o gráfico dos resíduos ponderados padronizados versus a ordem das observações. A partir dela é possível determinar se os resíduos apresentam algum padrão ou se estão distribuídos aleatoriamente em torno de zero, neste caso, observa-se que estão distribuídos de forma aleatória. Ainda é possível contabilizar a quantidade de observações fora do padrão estabelecido de [-2,2].

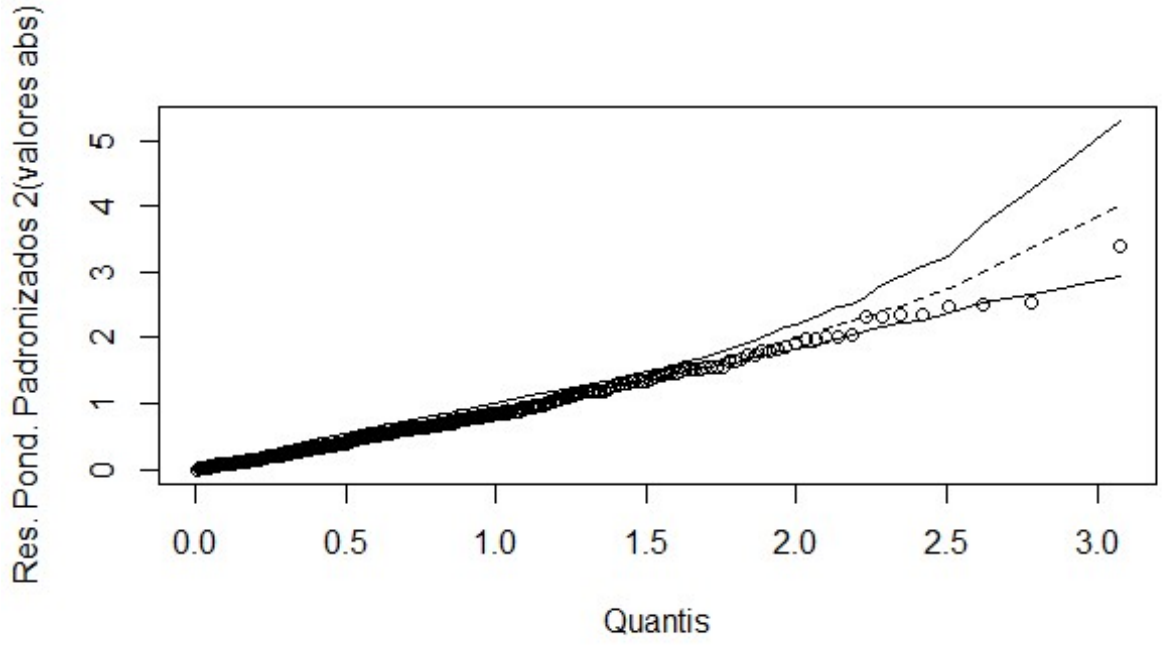

Figura 7: Gráfico da probabilidade Normal com envelope simulado

As medidas influentes auxiliam na investigação de possíveis observações discrepantes, por exemplo, os *outliers* que estão relacionados a variável resposta e o Leverage (ponto de alavanca) que está relacionado com as variáveis explicativas. Assim, é interessante identificar e verificar o impacto que cada uma dessas observações podem ocasionar nas estimativa dos parâmetros, visto que, podem influenciar na estimativa da reta de regressão.

Para quantificar a influência que uma observação exerce na estimativa dos parâmetros, (Espinheira, 2008) propuseram uma medida similar a Distância de Cook para os modelos

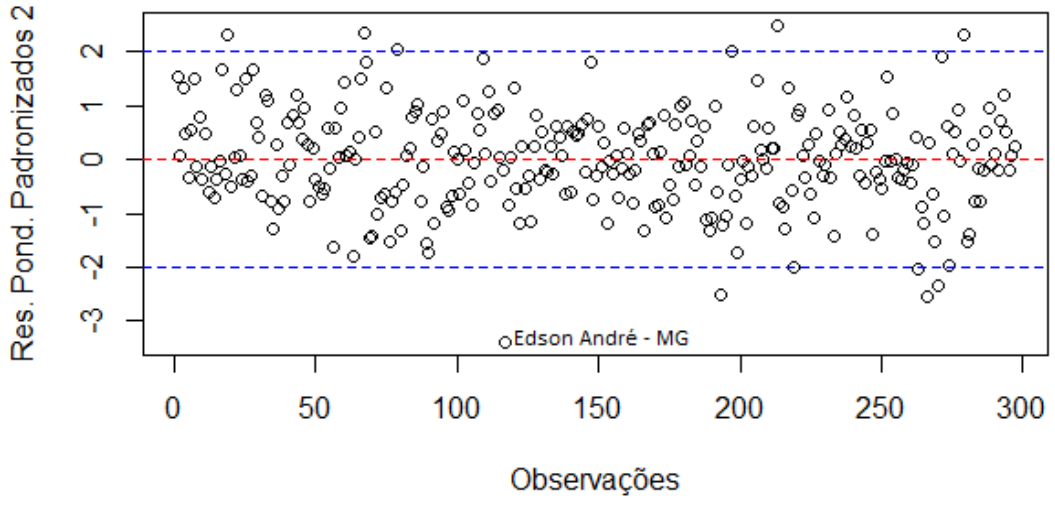

Figura 8: Resíduos Ponderados Padronizados 2 *versus* Índices das observações

de regressão beta. Na Figura 9, é observado vários pontos discrepantes. O interessante é que os candidatos que mais se destacam, são aqueles que obtiveram um percentual de votos alto tendo uma verba de campanha relativamente baixa (Soraya Thronicke(PSL), 16.19% de votos e R$76001.00 de receita; Marcos Val(PPS), 24.08% de votos e R$130930.00 de receita). Já quando medimos Leverage Generalizado, ver Figura 10, nota-se que os pontos que mais se destacam são os candidatos que tiveram um percentual de votos baixo tendo uma verba de campanha relativamente alta (Edson André(AVANTE), 0.017% de votos e R$360990.42 de receita;Aspasia(PSDB), 0.18% de votos e R$1030900.31 de receita).

A análise sem os 6 pontos influentes observados nas Figura 9 e Figura 10, resultou em um leve aumento nas estimativas dos parâmetros do modelo, com um R_p^2 de 0.7285. Porém, para este trabalho não foi retirado nenhum ponto influente como foi observado muita destas observações, não seria interessante retirar tantas observações. O mesmo acontece quando medimos o Leverage Generalizado, como pode ser visto na Figura 10.

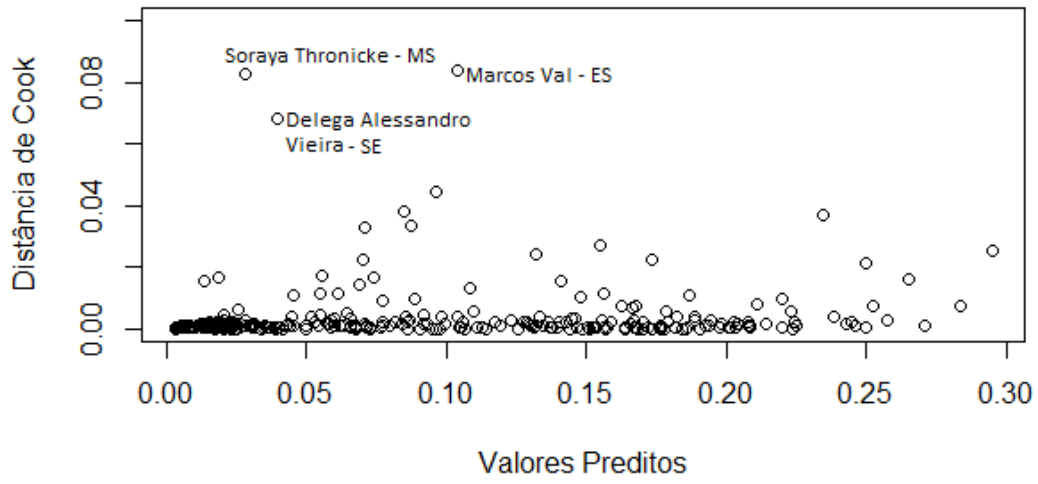

Figura 9: Gráfico da Distância de Cook

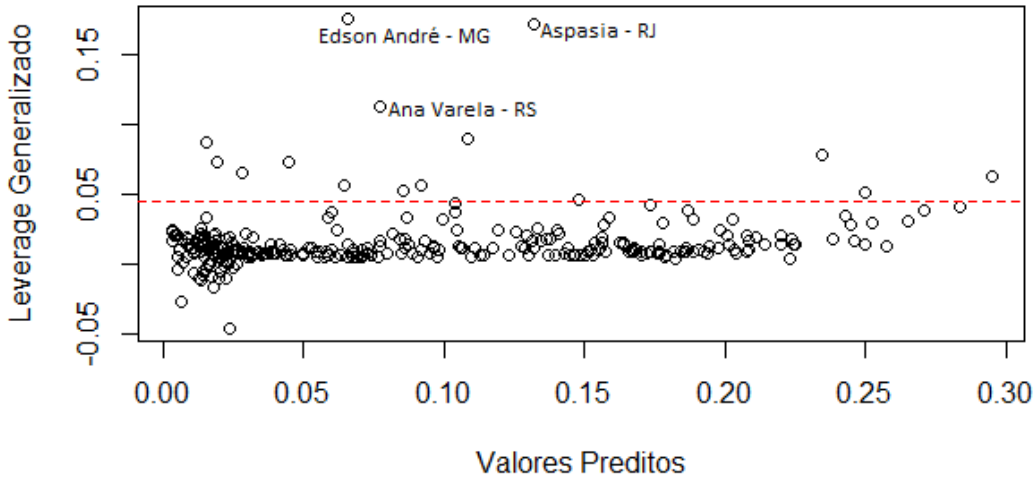

Figura 10: Gráfico do Leverage Generalizado

4.3.3 Ajuste dos Modelos de Regressão Múltipla Tradicionais

Para comparar a eficiência do modelo de regressão beta em relação aos modelos mais tradicionais, esta seção será destinada a comentar os possíveis problemas encontrados ao se ajustar modelos de regressão tradicional (com pressuposto de normalidade) no conjunto de dados proposto.

O modelo de Regressão Múltipla está disposto na Tabela 11. Repare que, assim como no modelo de regressão beta, a única variável não significativa para o modelo foi Reeleição. Após ajustar um modelo sem a variável Reeleição, foi obtido um AIC de -779.3737, consideravelmente menor que o obtido no modelo de regressão beta que foi de -1164,2640. Pode-se ainda citar mais dois problemas encontrados. O primeiro em relação ao Envelope Simulado, no gráfico da Figura 11, que o modelo apresenta desvio da distribuição de probabilidade proposta (Normal), o que já era esperada, uma vez que foi visto que a variável resposta não era simétrica. O outro problema encontrado, é que estão sendo preditos valores fora do domínio da variável resposta, o que já era esperado, visto que este tipo de ajuste não limita a variável resposta.

Tabela 11: Modelo de Regressão Linear Múltipla Tradicional

	Estimativa	Erro Padrão	t valor	Pr(>|t|)
Intercepto	-0.1583	0.0229	-6.92	0.0000
log(Receita total)	0.0163	0.0020	8.29	0.0000
Gênero:Masculino	0.0180	0.0099	1.82	0.0693
Reeleição	0.0068	0.0127	0.54	0.5926
Qt partidos coligação	0.0074	0.0011	6.96	0.0000

Tabela 12: Modelo de Regressão Linear Múltipla Tradicional, sem a variável Reeleição

	Estimativa	Erro Padrão	valor t	Pr(>|t|)
Intercepto	-0.1610	0.0223	-7.24	0.0000
log(Receita total)	0.0166	0.0019	8.70	0.0000
Gênero:Masculino	0.0183	0.0098	1.86	0.0643
Qt partidos coligação	0.0074	0.0011	7.01	0.0000

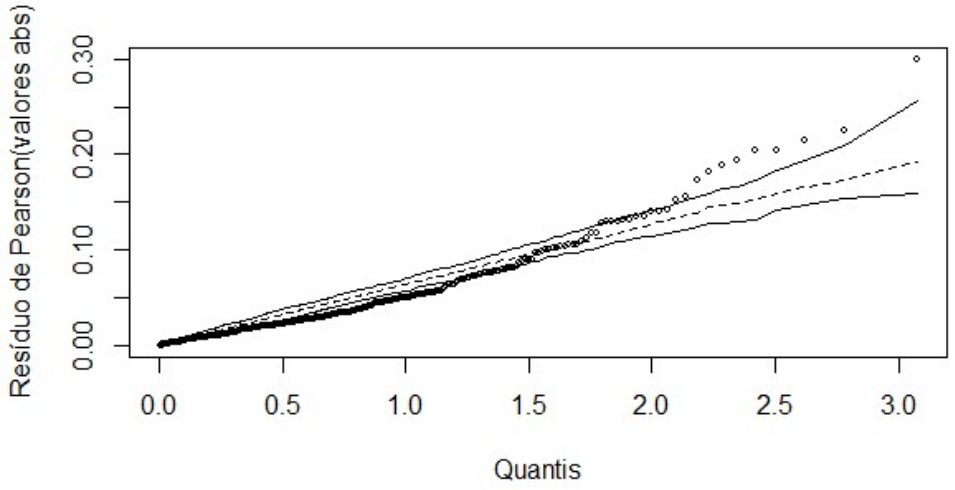

Figura 11: Gráfico da probabilidade Normal com envelope simulado

Uma das soluções apresentadas com relação ao domínio da variável resposta, seria aplicar uma transformação que expanda este domínio a todos os $\mathbb{R}$. A transformação sugerida é a Logito. O modelo ajustado com a transformação Logito na variável está apresentado na Tabela 14. Contudo, mesmo com a transformação na variável resposta, ainda foi verificado um comportamento heteroscedástico na variável resposta. Verifica-se na Figura 13 que há uma tendência, um pouco acentuada, da variabilidade dos resíduos aumentarem a medida que os valores preditos aumentam, chegando a um ponto em que esta variabilidade se mantem constante.

Por fim, como podemos notar na Tabela 15, apesar do modelo de regressão com transformação logito apresentar maior R_p^2, o AIC do modelo de regressão beta apresenta é consideravelmente menor que os demais. Por este, e o motivo da regressão beta poder ajustar modelos heteroscedásticos, ela se torna a melhor opção, entre estas apresentadas, para os ajustar um modelo que explique a natureza dos dados.

Tabela 13: Distribuição dos valores preditos do modelo de regressão tradicional

Min.	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máx.**
-0.0601	0.0347	0.0874	0.090	0.149	0.238

Tabela 14: Modelo de Regressão com Transformação Logito

	Estimativa	Erro Padrão	t valor	Pr(>|t|)
Intercepto	-9.7346	0.3212	-30.31	0.0000
log(Receita total)	0.4997	0.0275	18.17	0.0000
Gênero:Masculino	0.2558	0.1420	1.80	0.0725
Qt partidos na coligação	0.0604	0.0153	3.94	0.0001

Tabela 15: R_p^2 e AIC dos diferentes tipos de modelos estudados

Modelo	R_p^2	AIC
Modelo de Regressão Múltipla	0.5171	-779.3737
Modelo de Regressão Múltipla (c/ transf. Logit)	0.6970	-806.3772
Modelo de Regressão Beta com precisão variável	0.6968	-1164.264

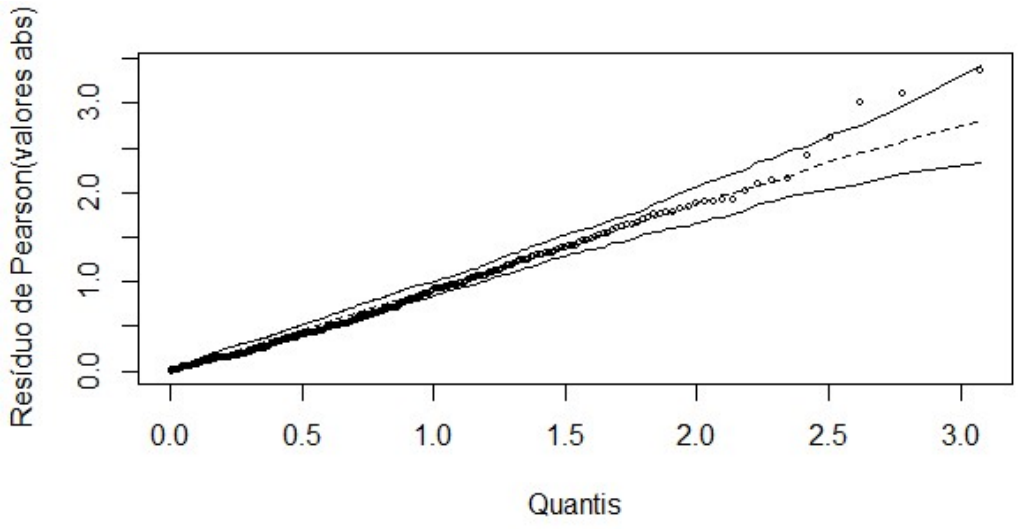

Figura 12: Gráfico da probabilidade Normal com envelope simulado do modelo com transformação logito

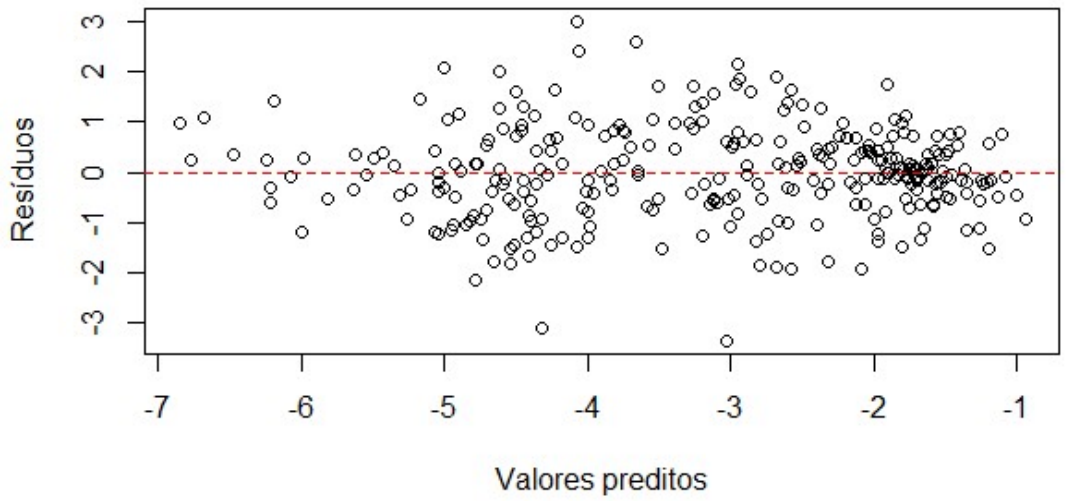

Figura 13: Gráfico de Resíduos versus Valores preditos

5 Considerações Finais

A classe dos modelos de regressão beta, na qual a variável resposta é modelada por meio de uma estrutura de regressão, é mais uma alternativa interessante para modelar variáveis que assumem valores limitados ao intervalo $(0, 1)$, podendo ser expandida para um intervalo (a, b), com $a < b$. De forma geral, o uso deste modelo acaba sendo mais apropriado do que os modelos mais usais, uma vez que ele apresenta uma flexibilidade ao se deparar com dados assimétricos e heteroscedásticos.

No ajuste do modelo de regressão beta, devido o fato da variável resposta apresentar um comportamento heteroscedástico, foi necessário ajustar um modelo para a precisão. Deste modo, foi verificado pela análise de diagnóstico que o modelo ajusta bem com os dados propostos.

No ajuste do modelo de regressão múltipla tradicionais, não foram atendido o pressuposto citados na Subseção 2.1. Além disso foram encontrados valores preditos fora do domínio da variável resposta. Este problema pode ser corrigido usando o modelo com transformação na variável reposta, e ainda houve uma correção na assimetria dos dados. Porém, os dados são tem uma característica heteroscedástica, que prejudicam o ajuste do modelo.

Verificou-se que os modelos de regressão beta para a modelagem do percentual de votos obtidos pelos candidatos ao Senado, apresentam um melhor ajuste, comparado aos modelos casuais de regressão. Pois, com ele é possível modelar tanto dados simétricos, como assimétricos, dados homoscedásticos e heteroscedásticos.

Por fim, o modelo de regressão beta proposto sugere que a quantidade de receita disponível e quantidade de partidos na coligação do candidato tendem a aumentar o percentual de votos, e que os candidatos homens tem em média uma chance de receber mais votos que as mulheres. Além disso, foi contatado, que a medida que a quantidade de receita aumenta a precisão do modelo diminui, portanto, não é fato que um candidato com muito recurso financeiro terá uma proporção de votos alta.

Referências

Andrade, A. C. G. d. (2007). *Efeitos da especificação incorreta da função de ligação no modelo de regressão beta*. PhD thesis, Universidade de São Paulo.

Atkinson, A. C. (1985). Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18.

Cordeiro, G. M. and Demetrio, C. G. (2010). *Modelos Lineares Generalizados e Extensões*. Piracicaba.

Cribari-Neto, F. and Zeileis, A. (2009). Beta regression in r.

Espinheira, P. L., Ferrari, S. L., and Cribari-Neto, F. (2008a). Influence diagnostics in beta regression. *Computational Statistics & Data Analysis*, 52(9):4417–4431.

Espinheira, P. L., Ferrari, S. L., and Cribari-Neto, F. (2008b). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7):799–815.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

Rocha, A. V. and Simas, A. B. (2011). Influence diagnostics in a general class of beta regression models. *Test*, 20(1):95–119.

Simas, A. B., Barreto-Souza, W., and Rocha, A. V. (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366.

Wei, B.-C., Hu, Y.-Q., and Fung, W.-K. (1998). Generalized leverage and its applications. *Scandinavian Journal of statistics*, 25(1):25–37.